Accepted Manuscript

Developing an ontology schema for enriching and linking digital media assets

Yoan Gutiérrez, David Tomás, Isabel Moreno

PII: S0167-739X(18)32385-9
DOI: https://doi.org/10.1016/j.future.2019.06.023
Reference: FUTURE 5041

To appear in: *Future Generation Computer Systems*

Received date : 4 November 2018
Revised date : 4 June 2019
Accepted date : 15 June 2019

Developing an Ontology Schema for Enriching and Linking Digital Media Assets

Yoan Gutiérrez[a,b,*], David Tomás[a,b], Isabel Moreno[a,b]

[a]University Institute for Computing Research, University of Alicante
[b]Department of Software and Computing Systems, University of Alicante
Carretera San Vicente del Raspeig s/n, 03690, Alicante (Spain)

Abstract

The abundance of digital media information coming from different sources, completely redefines approaches to media content production management and distribution for all contexts (i.e. technical, business and operational). Such content includes descriptive information (i.e. metadata) about an asset (e.g. a movie, song or game), as well as playable media (e.g. audio or video files). Metadata is organised following a variety of inconsistent structures and formats that are supplied by various content providers. Some challenges have been addressed in terms of standardising and enriching media assets metadata from a semantic perspective. Well known examples include Europeana and DBpedia. Nevertheless, due to the ongoing variability and evolution of digital contents, constant support and creation of new semantic representations are necessary. This article presents an ontology schema covering the requirements of users (content providers and content consumers) involved in the overall life cycle of a digital media asset, which has been designed and developed for a real scenario. The construction of this schema has been documented and evaluated following a methodology supported by quantitative and qualitative metrics. As part of the tangible results, the following outcomes were produced: (i) an RDF/XML schema available via Zenodo and GitHub; (ii) competence questions used for validation are published at GitHub; (iii) an exemplary ontology repository; and (iv) CRUD (Create, Read, Update and Delete) technologies for managing semantic repositories based on such schema. These results form an active part of the framework of a European project

*Corresponding author

Email addresses: ygutierrez@dlsi.ua.es (Yoan Gutiérrez), dtomas@dlsi.ua.es (David Tomás), imoreno@dlsi.ua.es (Isabel Moreno)

and other ongoing research initiatives.

Keywords:
semantic representation, ontology, digital media asset, entertainment industry

1. Introduction

Nowadays one of the challenges dealt by the new generation of technologies is managing the large quantity of existing information coming from digital media in different sources. This challenge requires a complete redefining of current approaches to manage multimedia content in all contexts: technical; commercial; and operational. These contents include descriptive information (i.e. metadata) about an asset (e.g. a movie, song or game), as well as playable media (e.g. an audio or video file). Additionally, descriptive metadata are used to describe and represent content by means of a variety of inconsistent structures and formats that are supplied by different content providers. Currently inconsistent data constitutes a bottleneck in the supply chain that can cause losses in sales [1].

Digital media content is usually represented by a broad range of specifications and vocabularies.[1] Such representations are mostly centred on industry actors (e.g. content providers and information brokers) and bibliographic data services, but not on viewers that consume media content anytime and anywhere, exclusively or in parallel with other activities. In this sense, the development of semantic-oriented standards, such as Europeana[2] and Dublin Core,[3] is becoming more frequent. This type of representation (i.e. semantic ontologies) allows the semantics to be captured behind a domain and provides intuitive metadata for representing and storing media content.

There is a variable sample of representation models that can support the research community. Most of them can be found in the Linking Open Data (LOD) Project index.[4] However, there is a lack of initiatives focused on covering the requirements of users involved in the overall life cycle of a digital media asset to include content providers as well as proactive and

[1] https://www.w3.org/2005/Incubator/mmsem/XGR-vocabularies/.
[2] http://www.europeana.eu/.
[3] http://dublincore.org/.
[4] http://linkeddata.org/.

interactive viewers. This deficit may result in the exclusion of important data that could benefit both content providers and viewers.

This paper deals with the above-mentioned deficit by presenting the Digital Media Asset (DMA) ontology schema, which has been developed as part of the EU-funded project "Socialising Around Media" (SAM)[5][6] (FP7-611312). Specifically, this work focuses on the design and development of the DMA ontology schema, representing digital media assets and capturing the requirements for industry and audience actors.

The goal of the SAM project is to build an advanced digital media delivery platform, combining second screen technologies[7] and content syndication in the domain of Social TV, where television and social media are united to promote communication and social interaction related to a broadcasted program content. The SAM ecosystem aims to collect a variety of characteristics around the entertainment scenario, which have also served as the basis for other related research projects.

The DMA ontology is directly involved in the fulfilment of different user requirements of the SAM project. The overall requirements of SAM that are supported by the use of the DMA ontology include: (i) offering additional, related, contextually relevant and personalised content to content providers for on-demand videos; (ii) assisting editors in the production and delivery of a wide range of second screen content experiences related to selected on-demand video and television programming. Additional content from the same broadcaster or third-party second screen content can be quickly referred to; and (iii) exposing and characterising (e.g. linking entities) from different sources in a unified asset description and data format, so that the asset can be found, used, linked and reused. These and other requirements are described in more detail in Section 4.2.

The main contributions made in this work are: (i) factorisation of terms and relationships (Section 4.2 and 4.3); (ii) reusability of shared ontology schemas (Section 4.3); (iii) formalisation and implementation of the DMA ontology schema (Section 4.4); (iv) detailed evaluation by considering different ontology metrics (Section 4.5); (v) testing and simulation of a real scenario (Section 4.5); (vi) accessibility to a permanent public documentation[8];

[5] `https://cordis.europa.eu/project/rcn/110682_en.html`, last access March 2019.

[6] `http://wiki.socialisingaroundmedia.com/`, last access March 2019.

[7] *Second screen* refers to a complementary electronic device, such as a smartphone, which allows the user to retrieve additional data about the content they are watching on the so-called *first screen*, usually a TV set.

[8] `https://w3id.org/media/dma/doc`, last access March 2019.

(vii) permanent access to the ontology schema [9] [10]; (viii) permanent access to an example scenario with instances described in this work [11]; and (ix) an API service for managing semantic repositories using Eclipse RDF4J[12] (Section 5).

The rest of the paper is organised as follows: Section 2 outlines the motivation for this research; Section 3 describes related works that serve as a guide and basis for the development of the ontology schema; Section 4 details the whole development process, as well as its evaluation, based on a published and freely accessible case study; Section 5 describes CRUD API services for managing semantic repositories based on the ontology schema proposed; Section 6 presents the market opportunities and comments on the previous case study; and finally, Section 7 concludes and proposes future work.

2. Motivation

Enormous amounts of data is processed every day and most of it is described in an appropriate format for their owners and consumers. In an effort to standardise the data interchange, XML, RDF/OWL, JSON-LD and other semantic descriptions have played a key role in the big task of providing not only information but also semantic content.

In order to create a reliable representation of the data in a specific domain, it is necessary to involve real data stakeholders in the standardisation process, which in the case of this work are Media Content Providers (MCP), i.e., providers of metadata for books, videos, DVD, games, etc. This type of actor would be interested in publishing their content assets on a marketplace platform, feeding data through a semantic driven transformation whereby no information is missed. These digital media assets, which have metadata and are semantically enriched, could be bought or used freely (depending on the business model) by other media providers to enhance their own assets. Platforms with this technology would facilitate media content providers to not only take advantage of their own content offers, but also to develop and share content by means of collaborative-based media enrichment mechanisms. This would be done by providing capabilities

[9] `https://w3id.org/media/dma`, last access March 2019.
[10] DOI: 10.5281/zenodo.2575666
[11] `https://w3id.org/media/dma/JamesBond_CaseStudy`, last access March 2019
[12] `http://rdf4j.org/`.

such as content processing, distribution through multi-format devices and channels, metadata extraction, secure storage, and efficient access/retrieval.

The potential users from the MCP, identified for this kind of ontology schema, are *editing staff*, *media data analysts* and *sellers*. The former would be interested in making use of this schema to populate it with new media metadata. These media metadata could be enriched by considering new semantic features and links to data from other MCP.

For instance, a MCP could use the DMA ontology schema to define information about the Danish actor *Mads Mikkelsen*, including data relating to awards received, social networks accounts, and representative keywords to facilitate the retrieval of this asset. Then, an editor could create an instance about the film *Casino Royale* and link it to the *Mads Mikkelsen* entry and other relevant sources of information, such as the Wikipedia entry to *Copenhagen* (his birthplace), different media covering the film (e.g. screenshots and trailers), and a DMA instance of the Ian Fleming's novel that inspired the film. In this way, the use of the DMA ontology facilitates the linkage with third-party data, creating richer contents by establishing semantic relations between the linked assets.

Regarding data analysts and sellers, they would be able to exploit the whole ontology for extracting innumerable combinations of semantic queries in order to generate reports based on concurrent evidences. At this stage, it would be possible to explore a large number of media assets imported and merged into existing semantic instances. To this end, a discovery process could be set up in which non-trivial semantic relationships would be revealed and, therefore, new enhanced data could be used for e-commerce. For instance, ternary relationships to discover connections between writers, whose work serves as inspiration of films, and the actors involved in those films.

Continuing with the previous example, a query of this type could retrieve actors that have participated in films based on novels written by Ian Fleming. An elaborated use case of the DMA ontology is presented in Section 6.1.

3. Related Work

There is a variety of well-known ontologies and semantic vocabularies with different goals that describe multimedia assets in a generic manner. The following paragraphs describe the most relevant initiatives for the SAM project. Readers interested in a thorough overview of multimedia ontologies can consult the work of Suarez-Figueroa [2].

ABC [3] is one of the first core ontologies dealing with the integration of multiple types of multimedia content (text, image, video, audio and web pages) for digital libraries, museums and the Internet. ABC is able to describe temporal entities (i.e. performances that happen to an object) and fundamental entities such as people, organisations and instruments.

Another relevant initiative is Simple Knowledge Organization System (SKOS) [4], which provides a common model for sharing and linking knowledge organisation systems, such as taxonomies or thesaurus, which share a similar structure for similar applications.

The Core Ontology for MultiMedia (COMM) [5] is proposed to represent semantically compound documents and their sub-parts (e.g. faces or objects detected in an image). COMM uses DOLCE [6] as its underlying foundational ontology.[13]

The Open Archives Initiative Object Reuse and Exchange (OAI-ORE) [7] defines a set of standards for the description and aggregation of web resources to optimally combine distributed resources from multiple media types.

Schema.org [8] is a collection of shared vocabularies that webmasters can use to mark-up HTML (HyperText Markup Language) pages to be understood by search engines.

Dublin Core Metadata Initiative (DCMI) Metadata Terms [9] is an RDF vocabulary designed to represent core features for resource descriptions (e.g. videos and images).

Multimedia Metadata Ontology (M3O) [10] was proposed to describe parts of documents in SMIL[14], SVG[15] or Flash[16] formats. More specifically, M3O aims at annotating rich, structured multimedia presentations that comprise diverse genres of content, such as images or text.

The European Data Model (EDM) [11] is an integration medium for collecting, connecting and enriching the descriptions from content providers. By organising cultural heritage assets digitalised throughout Europe, the EDM not only introduces new metadata, but also consolidates vocabularies such as ABC, DOLCE, DCMI, SKOS and OAI-ORE. Besides that, EDM can be extended with new elements to fulfil requirements from different but related scenarios.

[13] http://www.loa.istc.cnr.it/old/DOLCE.html
[14] https://www.w3.org/TR/2008/REC-SMIL3-20081201
[15] https://www.w3.org/Graphics/SVG
[16] https://www.adobe.com/devnet/f4v.html

In the LOD cloud there are also relevant initiatives for representing assets in the media domain.[17] Specifically, the BBC has developed three ontologies relevant for the SAM scenario: (i) BBC Ontology[18] describes products, web documents and platforms for which the BBC produces content; (ii) Programmes Ontology[19] aims to cover television content (i.e. brands, series and episodes), as well as the medium (i.e. channel, broadcaster and service), broadcast events and temporal annotations such as subtitles; and (iii) Music Ontology[20] provides main concepts and properties for describing music (i.e. artists, albums and tracks).

Other initiatives found from data-sets within the media LOD cloud, developed by other institutions, are: (i) Between Our Worlds ontology,[21] an initiative concerned with general metadata information about anime, including anime itself and its characters; (ii) LODE ontology,[22] provides a framework to describe historical events and their properties (e.g. the place where an event happened); (iii) Ontology for Media Resources,[23] developed by WC3, defines a core vocabulary (i.e. a set of annotation properties) for describing multimedia content available on the Web and mappings with existing vocabularies such as DCMI; and (iv) DBpedia Ontology,[24] a shallow, cross-domain ontology that has been manually created based on the info-boxes within Wikipedia such as film or book.

To date, these ontologies and schemas have focused on modelling knowledge from digital objects in terms of a wide range of metadata, but establishing few relations among their instances. The SAM project has several innovations that cannot be realised by exploiting these ontologies "off the shelf". The DMA ontology schema presented in this work is inspired in EDM, the ontology that consolidates more multimedia vocabularies. Hence, the DMA reuses, where possible, concepts and properties from EDM (which includes DCMI Metadata Terms and Schema.org) and DBpedia, but creates new ones to address specific requirements for media content annotation, linking and syndication.

Initiatives such as OAI-ORE combine objects to build aggregated re-

[17]https://lod-cloud.net/clouds/media-lod.svg.
[18]https://www.bbc.co.uk/ontologies/bbc
[19]https://www.bbc.co.uk/ontologies/po
[20]http://purl.org/ontology/mo/
[21]https://betweenourworlds.org/ontology/
[22]http://linkedevents.org/ontology/
[23]https://www.w3.org/TR/mediaont-10/
[24]http://mappings.dbpedia.org/server/ontology/classes/

sources like video and its metadata. The DMA schema extends this functionality by also allowing aggregated assets to be linked at a specific moment, which cannot be specified with the OAI-ORE model. Similarly, LODE ontology permits a descriptive account of when an event happened, whereas Programmes Ontology provides temporal annotations for subtitles; however, neither handles the second screen phenomena, since the assets that can be aggregated in DMA are products or agents that have to appear at a given time on a second screen.

A media asset representation has to consider not only how to represent knowledge (model), but also what format is necessary to store such information (representation). The latter refers to serialisation formats that allow the conversion of complex objects into sequences of bits. Many serialisation formats have been defined with the purpose of representing media data. For instance, JSON-LD [12] is a JSON-based format to serialise Linked Data information in a lightweight manner. Turtle (Terse RDF triple Language) [13] is a format that allows an RDF graph to be completely written in a compact and natural text form with abbreviations for common usage patterns and data types. OWL Manchester syntax [14] is a user-friendly compact and frame-based syntax for OWL ontologies. RDA (Resources Description and Access) [15] is a standard for resource description and access designed for the digital world that covers all types of content and media. RDF/XML [16] syntax describes how to encode RDF graphs in XML for data interchange on the web. RDFa (Resource Description Framework in Attributes) [17] provides a set of attribute-level extensions that can be used to embed metadata in HTML, XHTML and various XML-based document types. Similarly, Microformats [18] adds semantic mark-ups embedded and encoded within XHTML and HTML attributes. After exploring the above mentioned representation formats, JSON-LD was selected as the most appropriate one to store and manipulate assets within all SAM platform components. The rationale for this decision is explained in Section 4.4.

To sum up, the aim of the DMA ontology schema is to provide the required flexibility to describe effectively the various digital media entities and maintain compatibility with existing standards. This does not only refer to the possibility of integrating popular specifications in the DMA description, but also to mapping or linking assets if and when required (e.g. at a specific point in a video timeline).

4. Ontology Development

There are different methodologies available for building ontologies. For instance, BSDM [19] provides the guidelines developed by IBM for modelling companies as a preliminary step to developing Information Technology systems. The method proposed by [20] is one of the most comprehensive methodologies available for building ontologies. This method is described with an special emphasis in the capture phase, describing among others a procedure to identify the terms, produce definitions and handle ambiguous terms. KADS [21] proposes a structured way of developing knowledge-based expert systems. IDEF5 [22] is a software engineering method to develop and maintain usable, accurate, domain ontologies. Tom Gruber's principles for ontology design [23] provides an engineering perspective on the ontology development. On-To-Knowledge [24] proposes building ontologies by taking into account both knowledge process (i.e. handling knowledge items) and knowledge meta-process (i.e. introducing and maintaining knowledge management systems) capture. DILIGENT [25] provides a methodology for collaborative and distributed ontology engineering. Ontology Design Patterns, also known as *pattern-oriented*, is a modelling approach in which reusable templates (patterns) are defined to encode knowledge. This methodology is commonly used to design ontologies in the media domain. For example, [10] defined patterns to convert MPEG-7[25] standard to an ontology that is used to generate multimedia presentations. More recently, an scenario-oriented methodology called NeOn has been proposed [26]. This approach suggests a variety of pathways for developing ontologies and ontology networks. NeOn has been successfully applied in the multimedia domain [27].

In this work, METHONTOLOGY [28] has been chosen from the existing literature as the most suitable for developing task-oriented ontologies, and one of the most comprehensive ontology engineering methodologies. METHONTOLOGY enables the construction of ontologies from the knowledge level (i.e. the conceptual level) to the implementation level, and proposes a development life cycle, techniques, outcomes, and evaluation principles for implementing ontologies.

The development life cycle of the DMA ontology schema was carried

[25] https://mpeg.chiariglione.org/standards/mpeg-7

out with Protégé,[26] Protégé visual plugins (OWL Viz,[27] OntoGraf,[28] and VOWL[29] among others) and query language tools (e.g. SPARQL and DL Query) [29]. A solid foundation for this research was provided by reworking and restudying media content scenarios. Major efforts were focused on reusing existing ontology terms that could provide standard and stable descriptions to media entertainment ecosystems.

In the next sections, a detailed description of the main stages carried out for the ontology development life cycle is explained: plan; specification; conceptualisation; reuse and integration; formalisation and implementation; and evaluation and control.

4.1. Plan

All the tasks involved in the planning process are properly documented in the SAM project reports and in scientific papers [30][31]. The planning included two iterations following the steps described in sections 4.2 to 4.5. At the end of the first iteration, which took ten months, a first draft was obtained and used by the SAM project partners to test their components. After that, the goal was to collect during the next five months their feedback, running a second and final iteration.

Three main tasks were carried out during the planning phase: (i) the study of the state-of-the-art in semantic representation; (ii) the analysis of the requirements from user stories and scenarios; and (iii) the examination of component dependencies for describing the DMA ontology schema. To illustrate the context in which this schema was conceived, Figure 1 presents the SAM's infrastructure and components that inspired its design and development. The components directly interacting with the DMA schema, coloured in grey in the aforementioned figure, are as follows: *Content Gateways*; *Semantic Services Linker*; *Marketplace*; *Brand and Consumer Protection*; and *Syndicator*.

Fully understanding the context in which the DMA ontology schema has been conceived requires an awareness of the SAM project goals, which were mentioned in the introduction section.

[26] `http://protege.stanford.edu/`.
[27] `http://protegewiki.stanford.edu/wiki/OWLViz`.
[28] `http://protegewiki.stanford.edu/wiki/OntoGraf`.
[29] `http://protegewiki.stanford.edu/wiki/VOWL`.

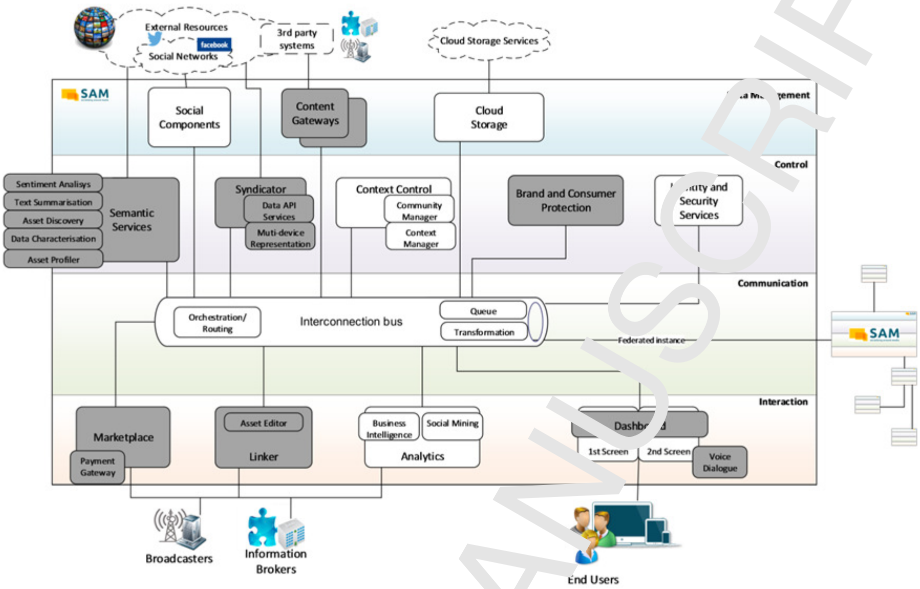

Figure 1: SAM framework for exploiting semantic assets

4.2. *Specification*

This phase consisted of constructing an ontology schema specification document written in natural language, using a set of intermediate representations or using competency questions.

Taking into consideration the SAM scenario, the following elements have been specified: (i) a set of 28 competency questions obtained from end users' commercial metadata (see Section 5); and (ii) a list of 167 user requirements collected by 19 members of the team. In this respect, Table 1 presents a description of the partners of the SAM project involved in the specification of the system requirements. Table 2 shows five examples from the list of user requirements gathered.

4.3. *Knowledge Acquisition, Conceptualisation, Reuse and Integration*

This phase involved data cross-checking performed through 13 interviews that included users, stakeholders and the team in charge of the SAM components. Three brainstorming sessions involving the project team also took place. Seven major iterations (involving a change in the version number of the DMA schema) were required to refine the terms extracted and for structuring the domain knowledge in a conceptual model. Each major iteration involved an average of nine minor iterations (involving small

Table 1: Partners of the SAM project involved in the specification of the user requirements

Company/University	Expertise	Role	Individuals
TIE Nederland B.V. (The Netherlands)	Software development	Technology provider	2
Ascora GmbH (Germany)	Software development	Technology provider	2
Talkamatic AB (Sweden)	Voice recognition and interactive systems	Technology provider	3
TP Vision Belgium NV (Belgium)	First and second screen applications	Technology provider	1
National Technical University of Athens (Greece)	Distributed systems and service oriented architecture	Research	3
University of Reading (UK)	Intelligent systems and machine learning techniques	Research	1
University of Alicante (Spain)	Information systems and ontology development	Research	3
Deutsche Welle (Germany)	Broadcasting technologies	Content provider	2
Bibliographic Data Services Limited (UK)	Content management and metadata creation	Content provider	2

changes that added to consolidate a new version of the schema). This process implied determining concepts, attributes, relationships and restrictions. Finally, from the three latest major iterations, vocabularies from other ontologies (e.g. schema.org) were reused and integrated. This process was performed by identifying terms that were semantically coherent with the terms identified in the conceptualisation. In each major iteration, a version file was created until the final version 1.0 was attained. During the development process, the average agreement among team members was around 95%. Terms, relations and concepts not agreed on (i.e. around a 5%), were not included in the final version.

Once refined, the terms extracted formed a glossary of terms. Subsequently, a conceptual map[30] was designed, which served as a guideline for identifying key aspects that directly affected the ontology schema specification.

Based on the SAM scenario, the conceptualisation presented in Figure 2 was designed. This figure represents different semantic relationships inherited at three levels in the DMA ontology schema. First, the main class Asset provides basic relations to state the ownership (also the creator

[30] http://wiki.socialisingaroundmedia.com/index.php/Reference_Model, last accessed March 2019

Table 2: Examples from the set of 167 user requirements collected. Possible values of *Priority* column are mandatory (*M*), desirable (*D*), optional (*O*) and exceptional (*E*)

Priority	User stories	Primary type of interaction	Secondary type of interaction	Primary target group
D	As a TV user, I can explore content through semantic descriptions that make sense, so that I can actively search for interesting content.	Search and discovery	Data linking	End users
M	As an end user I get content through the SAM platform that has filtered out any links to illegal sites (e.g. pirate copies of films, music, games).	Rights management and protection	Control	End users
D	As a TV user, I can receive content that has been automatically found for me on the web, social media and media markets.	Search and discovery		End users
D	As an end user I know I am receiving information about the correct person or asset-based on unique identifiers.	Search and discovery	Rights management and protection	End users
M	As an experience provider, I have access to efficient and effective context-sensitive linking: only related content is linked (e.g. Film "Casino" is not linked to ads on casinos but only to relevant material).	Content syndication	Data linking	Content providers

and publisher), generic relations with other assets, and external sources like DBpedia. The second level presents *Organisation*, *Person* and *Product*, ensuring contribution and authorship relationships between them. Finally, a third level specifies the relationship between specific products, such as albums, films, games and music recordings.

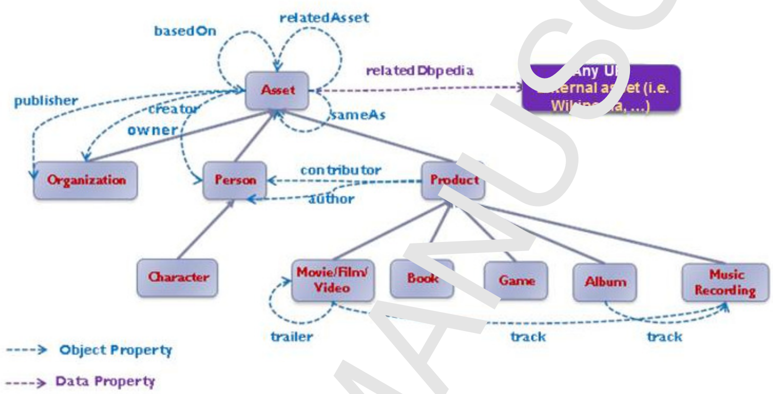

Figure 2: Main relationships in the DMA ontology schema

Once these basic relationships among the different asset concepts defined were considered, the next step consisted of determining the specific properties that characterise them. A brainstorming was performed to collect all potentially relevant terms, and sample data from a content provider partner was also examined to identify the key concepts and relationships in the SAM domain. At this stage, the terms represented concepts of the ontology schema. A grouping operation by using *part-of* terms' organisation was performed to initially categorise the terms for inclusion, exclusion or borderline. This resulted in a list of terms related to each media entity managed in the SAM project, revealing potential class attributes or relationships.

After this initial grouping, common terms were detected (e.g. *name* and *title* attributes occur in different entities). At the same time, unambiguous text definitions for such concepts and relationships were accurately identified from the SAM scenario. This strategy avoided misunderstanding among concept names and relationships, providing a higher level of abstraction. These generic concepts were divided into the three main categories shown in Figure 3, which were considered the starting point to place the remaining concepts. These concepts capture the knowledge about three main issues that are explicitly stated in the SAM scenario: (i) model the different asset classes; (ii) define nomenclatures such as language, country and others;and finally, (iii) model the extension concepts that represent the information needed by SAM components for capturing operational behaviour. In summary, the DMA ontology schema describes the media content by using three main classes: *Asset*; *Nomenclature*; and *Extension*.

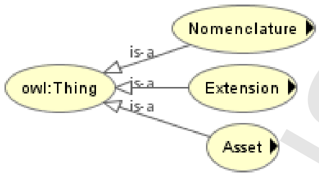

Figure 3: Root concepts

Asset represents major concepts in this work, such as *Organisation*, *Person* and *Products* (*Book*, *Movie*, *Video Game*, *Music Album* and *Music Recording*). Figure 4 illustrates the semantic relationship between the digital media assets, which are inherited at three levels.

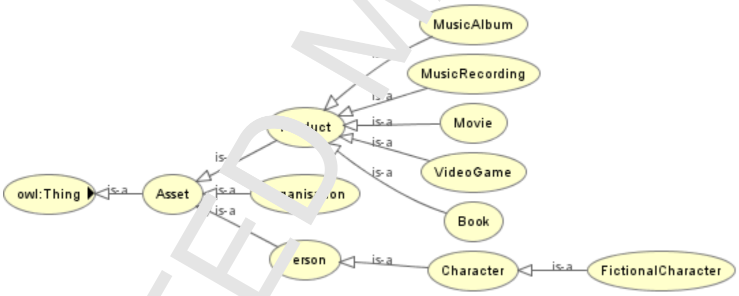

Figure 4: Asset classes

The second class, *Nomenclature* (see Figure 5), represents the closed sets of values that can be assigned to different parameters defined in the ontology schema, such as *Profession* (e.g. *director* or *writer*) or *Language* (e.g. *en*, *fr*, *de*, *it* codes).

The third class, *Extension* (see Figure 6), involves all the specific extensions needed to set particular edition features required by assets into a multimedia platform:

- *Voice Control Extension* represents a specific lexicon or grammar used in the platform for speech recognition tasks. In this way, the platform

Figure 5: Nomenclature classes

recognises spoken words associated to an asset returning a response related to it.

- *Semantic Extension* represents semantic relationships between assets without considering a specific type of relationship, like *owner*, *contributor*, *author* or *track*. Another relationship that can be established involves assets and DBpedia entries. Figure 13 shows an example.

- *Generic Extension* stores key-value pairs of elements that represent information that is not covered by other concepts in the ontology schema.

- *Social Media Extension* represents information about social media sources related to an asset. For example, the official Facebook page of Casino Royale could be included as part of the asset information related to this film

- *Postal Address Extension* (reused from Schema.org[31] by assigning an

[31] `https://schema.org/PostalAddress`.

alternative label) includes information about the postal address associated to an asset.

- *Syndication Extension* includes the information required to syndicate an asset and deploy it in a second screen environment.
- *Linking Extension* represents timelines associated to a specific asset for the purpose of visualisation on screen. Such is the case of subtitles in films, which are shown at a specific point and for a fixed time span.
- *Owner Extension* provides generic information about the owner of the asset, including fields for representing data by using key-value pairs. This representation allows setting any data not yet defined in the ontology.

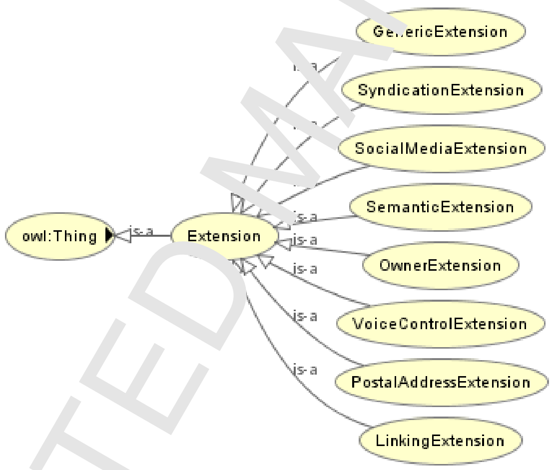

Figure 6: Extensions classes

After defining these three root concepts, the process continued identifying terms to define such concepts and relationships, producing and completing all the definitions in the SAM scenario. In this step, hypernym relationships (*is-a*) were generated by analysing the connection between terms and hierarchy levels. Different types of *part-of* relationships were generated (i.e. object properties and data properties) as described in Figure 7. These relationships were semantically organised in a hierarchy for a better understanding. Additional decisions were made in this stage, since different terms seemed to correspond with the same concept definition (e.g.

title and *name*). In these cases, one of them was chosen as representative of these conflicting terms, as suggested in [28]. For instance, in the previous example *title* was chosen instead of *name*.

Asset Classes	Properties		
(A) dma:Asset /(O) schema:Organisation; (P) schema:Product; (B) schema:Book; (G) schema:VideoGame; (M) schema:Movie /Video/Film; (MA) schema:MusicAlbum; (R) schema:MusicRecording	rdf:id	schema:duration	schema:awards
	rdf:type	dbo:price*	dma:creationDate*
	schema:url*	**dma:owner***	dma:identifier
	schema:image	**schema:contentRating***{schema:requiredMinAge*}	dma:lastUpdated*
	dc:language	schema:ratingValue*	dma:screenshoots
	schema:addressCountry*	schema:title*	schema:alternateName
	dma:sourceValue*	**schema:status***	[illegible]:copyright
	dma:syndicationExtension [**dma:theme***, **dma:widget***]	**dma:socialMediaExtension**[dc:description*, dma:identifier, schema:image, schema:url* **dma:sourceType***]	**dma:voiceControlExtension**{ dma:grammar}
	dc:description*	**dma:linkingExtension**[dma:targetTitle*, dma:targetType*, dma:linkStart*, dma:linkEnd*, schema:validUntil*, dma:sourceElement dma:targetElement *, dma:targetId*, **dma:owner***, **dma:theme***, **dma:widget**]	**dma:semanticExtension** { **dma:relatedAsset**, schema:keyword, dma:relatedDbpedia }
	schema:genre		
	schema:sameAs		
	dc:creator*		**dma:genericExtension**[key*,value*]
	schema:license*		
	schema:validFrom*		**dma:ownerExtension**[key*, value*]
	schema:validUntil*		
P	**dbo:basedOn**	**schema:publisher***	dma:referenceNumber
	dma:barcode*	dma:exVatPrice*	dma:vatPrice*
	dbp:adaptation	**schema:auth**[illegible]	schema:typicalAgeRange*
	dma:catalogueNumber*	**dma:pubCount**[illegible]	dcTerms:issued*
	dma:edition*	**schema:contributor**[[illegible]:role+ { **owl:annotatedSource+** , **[illegible]:annotatedTarget+**}]	**schema:encodingFormat***
B	**schema:author**	**schema:add**[illegible] { schema:addressLocality*, [illegible]ma:email, [illegible]ema:faxNumber , [illegible]stalCo[illegible] , schema:streetAddress *, sche[illegible]:tele[illegible]one, **schema:addressCountry** }	dbo:notes
	schema:isbn*		schema:numberOfPages*
	dma:lc		schema:reviewBody
	dma:bic		purl:series
	schema:commentText		dma:seriesPart*
	dma:deway		schema:text*
G	dma:accessories	[illegible]hema:gamePlatform	dma:numPlayers*
M	dma:digitalIdentifier*	dma:[illegible]umber*	purl:series
	schema:duration*	schema:numberOfItems*	**schema:subtitleLanguage**
	schema:track	**schema:trailer**	purl:series
MA	schema:duration	dbo:notes	**schema:track**
	schema:numberOfItems	schema:numTracks*	dma:bic
R	schema:position[illegible]	**dma:recording$**	schema:duration
	dma:trackISR[illegible]		
(PC) schema:Person / dma:Character	**schema:affi[illegible]tion**	**schema:deathPlace***	**schema:nationality**
	schema:bi[illegible]Place*	schema:familyName*	dbo:notes
	rdvocab:dateO[illegible]	**rdvocab:gender***	**rdvocab:professionOrOccupation**
	rdvocab[illegible]ateOfDea[illegible]	schema:givenName*	

Figure 7: Term and class distribution. In **bold** those considered as object properties; * represents functional properties with cardinality 1x1; $ represents symmetric relationships; and + represents annotations inside relationships. The context abbreviation "dma" refers to the proposed schema (https://w3id.org/media/dma#).

Note that in Figure 7 each term includes a prefix, in accordance with the namespaces listed in Table 3. This namespace allows identifying the original source from which a term is created or reused. A complete description of each term is available in the published ontology schema and documentation web page referred at the beginning of Section 4.

After several revisions analysing and discussing the previous issues and the two aforementioned iterations, the schema of the DMA ontology was obtained. This schema comprises a list of terms that were reorganised as

Table 3: Reused external sources in the DMA ontology schema

External Sources	Abbreviation	Namespace
ONTOLegolang	ONTOLegolang_UAge	https://w3id.org/nlp/ONTOLegolang_UAge#
Schema.org	schema	https://schema.org/
RDA Registry	rdvocab	http://rdvocab.info/ElementsGr2/
OWL Schema	owl	http://www.w3.org/2002/07/owl#
SWRL	swrlb	http://www.w3.org/2003/11/swrlb#
Protégé	protege	http://protege.stanford.edu/plugins/owl/protege#
XML Schema	xsd	http://www.w3.org/2001/XMLSchema#
Europeana	edm	http://www.europeana.eu/schemas/edm#
RDF Syntax	rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
DBpedia	dbp	http://dbpedia.org/property/
DBpedia	dbo	http://dbpedia.org/ontology/
Dublin Core	dc	http://purl.org/dc/elements/1.1/
Dublin Core	dcTerms	http://purl.org/dc/terms/
RDF Schema	rdfs	http://www.w3.org/2000/01/rdf-schema/

concepts, relationships and attributes.

One of the goals of the DMA schema is to integrate and reuse external ontology sources. For this reason, during the capturing and coding process, questions arose about when and how to reuse already existing ontology schemas. The DMA ontology schema was developed according to the semantic information defined in Europeana, which in turn is mostly based on DCMI. The rest of the reused terms were borrowed from Schema.org and DBpedia. This reusability (see Figure 7) allows representing metadata and linked data of different simple and generic resource descriptions. Most of the metadata was aligned to external sources of the semantic web to achieve the Level 3 of interoperability[32] that these shared sources propose [32] (see Table 3).

As shown in Figure 7, a wide variety of relationships were considered for connecting instances. They range from the most semantically descriptive such as *language*, *addressCountry*, *genre*, *sameAs* or *creator*, to the more generic, like *relatedAsset* and *relatedDBpedia*.

A special case occurs when assigning roles for representing a contribution relationship between an *Asset* and a *Person* or *Organisation*. For example, as shown in Figure 8, "Daniel Craig is a contributor to Casino Royale" and not "Casino Royale is a contributor to Daniel Craig".

[32] Semantic Interoperability means "a common information exchange reference model is used, allowing the meaning of the data to be shared".

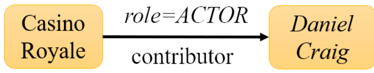

Figure 8: Contributor annotation solution

Hence, two annotations were considered to ensure the proper orientation of the contributor's role (*dbp:role*): *owl:annotatedSource*, which determines the subject of an annotated axiom, and *owl:annotatedTarget*, which determines the object of an annotated axiom. As a result, the example can be modelled as shown below.

Listing 1: Example of a the contributor relationship in JSON-LD format. JSON attributes are highlighted in bold

```
schema:contributor{
    dbp:role = "ACTOR"
    owl:annotatedTarget = "Daniel_Craig"
    owl:annotatedSource= "Casino_Royale"}
```

4.4. Formalisation and Implementation

The goal of this phase is two-fold. On the one hand, the ontology schema (model) needs to be defined using a formal language that supports the ontology representations selected at the previous integration phase. Therefore, an explicit and formal representation of the conceptualisation model was defined (see Figure 9). Supported by Protégé Desktop 5.0, every concept was transformed into an ontology *Class*, every relationship into an *Object Property*, and every attribute into a *Data Property*. In the life cycle of this phase, the ontology schema restrictions were stored in an OWL file. The structure of the schema can be explored using the WebVOWL interface.[33]

On the other hand, instances (i.e. assets) must be stored and manipulated by all SAM platform components using a serialization format suitable for representing media data in a lightweight manner. To this end, a previous work carried out by the authors [33] analysed the implications of adopting the serialisation formats, previously described in Section 3, for every component of the SAM platform. The conclusion was that JSON-LD[34] was the most appropriate format to store and manipulate assets for several reasons:

[33] http://www.visualdataweb.de/webvowl/#iri=https://w3id.org/media/dma, last access March 2019.

[34] https://www.w3.org/TR/json-ld/.

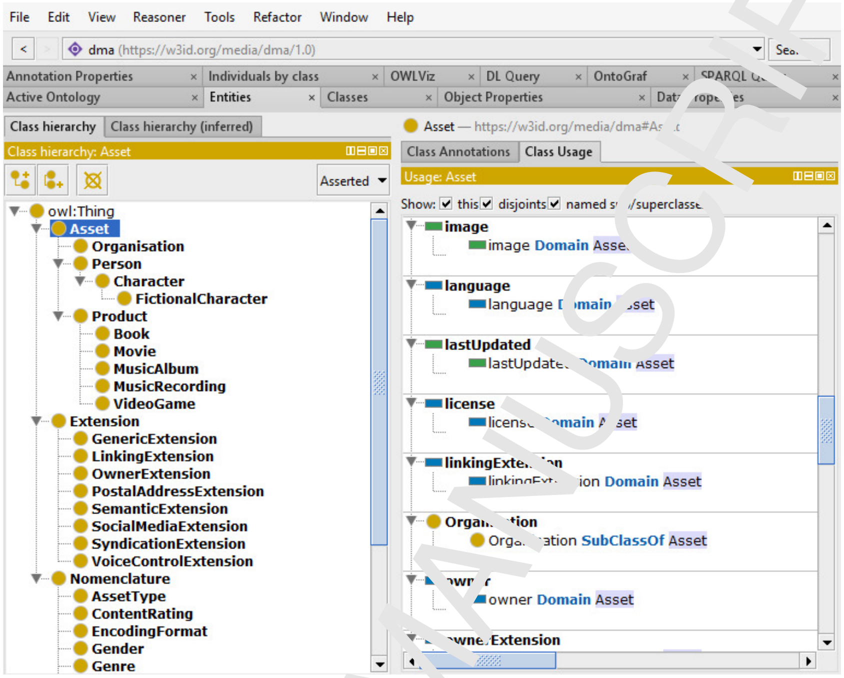

Figure 9: Designing the schema

(i) JSON-LD is a JSON-based format to serialise Linked Data; (ii) the syntax is designed to facilitate the integration into deployed systems that already use JSON; and (iii) JSON-LD provides a smooth upgrade path from JSON to JSON-LD, since traditional JSON can be transformed to its semantic counterpart. For all the reasons mentioned above, JSON-LD was chosen as the format to facilitate the development of the SAM platform.

4.5. Evaluation and Control

The means by which ontologies are evaluated entails an assessment of the quality of the final representation, in terms of maintenance and reusability. *Quality* is understood as the degree to which a set of functional and physical characteristics matches the needs and expectations established in the specification phase [34]. Unfortunately, there is a disagreement on the way qualitative and quantitative validations are carried out [35] [36] [37] [38] [39].

The current trend is to accept that the main purpose of an evaluation is to check that the conceptualisation model matches the adequacy of its content (validation) to determine their usefulness and potential for reusing.

The aim of validating the ontology schema consists of ensuring that there are no construction errors or defects. In addition, this schema is verified matching its definitions, as close as possible, to the domain for which the schema was created.

The following sections describe the set of qualitative and quantitative features that have been used to validate the DMA ontology schema.

4.5.1. Qualitative Validation

As mentioned before, the DMA ontology schema describes the media content using three main classes: *Asset* (containing 10 subclasses); *Extension* (including 8 subclasses); and *Nomenclature* (comprising 14 subclasses). Each subclass may have additional properties added to the properties inherited from their parent classes.

The ontology schema proposed does not contain any loop issues in the hierarchical structure modelled (i.e., it does not have any class defined as a generalisation and specialisation of itself). The hierarchical relationships between subclasses is transitive (if *B* is a subclass of *A* and *C* is a subclass of *B*, then *C* is a subclass of *A*) and all the declared sibling classes in the hierarchy are at the same level of granularity (see Figure 4, Figure 5 and Figure 6). Also highlight that a functional property sets the cardinality for a single value (1:1), which establishes a single and not multiple values for these properties. For example, a *Movie* has a single *title*, not a list of titles, and an *Asset* has a single *Owner*, not multiple owners.

Three different types of active relationships were considered to improve the expressivity of the DMA ontology schema, all of them with their respective cardinality: hierarchical relationships between assets (*rdf:subClassOf*); relationships between assets and classes used as nomenclatures (e.g. and *dc:language*, *schema:affiliation* and others); and relationships for representing particular extensions of the assets necessary in a multimedia platform like SAM (see Section 4.3). Furthermore, this ontology schema was tested applying the standard W3C validator,[35] obtaining a successful result.

4.5.2. Quantitative Validation

Some metrics proposed in [36] [37] [38] [39] were selected and divided into two groups to determine the physical characteristics of the structure and the type of content described in the ontology schema. The first group of

[35] `https://www.w3.org/RDF/Validator`, last access March 2019.

metrics, shown in Table 4, refers to descriptive metrics, whereas the second group, in Table 5, provides average metrics.

In Table 4, (1) refers to the sum of class axioms (*SubClassOf* counts) and (2) refers to the sum of object properties (*Transitive*, *Inverse*, *Functional*, *SubPropertyOf*, *Symmetric*, etc.).

Table 4: Protégé and manually calculated metrics

Metric	Result
Protégé	
Axiom	1066
Logical Axiom Count	558
Class Count	35
Object Property Count	49
Data Property Count	93
Annotation Terms Count	311
Equivalent Class	0
Sub-Object Property	86
Functional Object Property	18
Equivalent Object Property	0
Symmetric Object Property	1
Sub-Data Property	124
Functional Data Property	51
Equivalent Data Property	5
Manually calculated	
Taxonomic Relationships N. (1)	32
Other Non-Taxonomic rel. (2)	49
Root Classes N.	3
Intermediary Classes N.	4
Leaf Classes N.	28
Equivalent Relationships N.	4
Reused Classes	13
Reused Object Properties	24
Reused Data Properties	40

The results in this table show that the ontology schema presents a notable quantity of axioms which are distributed into different categories, i.e., classes, properties, restrictions, etc. Based on the counts presented in the *Result* column, different formulas were evaluated with the aim of quantitatively validating the DMA ontology schema.

The results of these formulas are presented in Table 5. In the formula labelled as (3), s represents the number of direct subclasses of a concept i, while c is the total number of concepts; in (4), r describes the total number of taxonomic relationships of a concept i; in (5), *r_noT* represents the number of non-taxonomic relationships of a concept i; in (6), *reused_rel(i)* describes the number of reused terms (e.g. DCMI) of a concept i; in (7), *reused_prop(i)* represents the number of reused attributes of a concept i; in (8), *path(i)* describes the deepest path from a concept i to leaf node; in (9), *n_att(i)* represents the number of data properties/attributes of a concept i and *n_rel(i)* the number of object properties of that concept; in (10), *sc(i)* describes the number of subclasses of a concept i; finally, in (11), *tax_rel(i)* represents the number of taxonomic relationships of a concept i and *sem_rel(i)* the number of non-taxonomic (semantic) relationships.

Table 5: Formulas for metrics. Column *Result* shows the average and the interval with minimum and maximum values

Metric	Formula	Result
Subclasses (Avg.subclasses.n)	$\frac{\sum_{i=1}^{c} s(i)}{c}$, (3)	0.91 [0,13]
Taxonomic rel. by class (Avg.rel.n)	$\frac{\sum_{i=1}^{c} r(i)}{c}$, (4)	0.91 [0,13]
Non-taxonomic rel. by class (Avg.notrel.n)	$\frac{\sum_{i=1}^{c} r\_not(i)}{c}$, (5)	1.22 [0,35]
Semantic reused rel. by class (Avg.reuse rel)	$\frac{\sum_{i=1}^{c} reused\_rel(i)}{c}$, (6)	0.74 [0,8]
Reused attributes by class (Avg.reuse prop)	$\frac{\sum_{i=1}^{c} reused\_prop(i)}{c}$, (7)	1.48 [0,9]
Avg. depth of inheritance by class (Avg.depth)	$\frac{\sum_{i=1}^{c} max(path(i))}{c}$, (8)	0.18 [0,2]
Property density (Prop.density)	$\frac{\sum_{i=1}^{c} n\_att(i)+n\_rel(i)}{c}$, (9)	3.77 [0,45]
Inheritance density (Inh.density)	$\frac{\sum_{i=1}^{c} sc(i)}{c}$, (10)	1.17 [0,14]
Relationship density (Rel.density)	$\frac{\sum_{i=1}^{c} tax\_rel(i)+sem\_rel(i)}{c}$, (11)	2.14 [0,21]

The average metrics presented in Table 5 lead to several conclusions. This ontology schema has an appropriate and balanced weight in both vertical and horizontal axis of the inheritance tree. However, taking into account the value 1.17 in the inheritance density parameter, this ontology schema can be classified as domain specific, as suggested in [36], compared to more generic ontologies such as SWETO [36] (with an inheritance density of 4.00) or TAP [40] (with 5.36). Ontologies with low inheritance density has a prevalence of the vertical axis, which may reflect a more specific type of knowledge representation, whereas in ontologies with high inheritance density the horizontal axis dominates, representing more general knowledge [36]. The relative low density (0.91 average subclasses by concept) illustrates the restrictions mentioned in Figure 7: functional properties with cardinality 1x1; symmetric relationships; and annotations inside relationships. These limitations also explain the cohesion achieved (0.18 average depth of inheritance), which is very close to an average level [35][36]. However, the main advantage of this ontology schema lies in the completeness of relations and declared properties. In this respect, non-taxonomic relationship density (1.22 on average by concept) shows its potential for addressing semantic inferences (2.14 relationship density) [39], as well as for reusing it for future ontology alignments (average of relationships reused by concept 0.74 and reused attributes 1.48) [34][37]. Finally, the knowledge density is solid, as the ontology schema is extensive and detailed (3.77 property density). This makes population with either low or high density data easier, in a manual or automatic way [34][37].

4.5.3. Validation of Competence Questions

A set of 28 competence questions was designed by the stakeholder experts mentioned in Table 1. The aim of these questions was to determine whether an ontology repository, based on the DMA schema, could provide a correct response to these questions, validating the correctness of the ontology in its context of use.

These competence questions had different degrees of difficulty, ranging from simple questions (e.g."Which films starring Mads Mikkelsen are in English?") to more complex questions (e.g. "Which films feature collaborations between director Martin Campbell and actor Daniel Craig?" or "Which authors have written books based on James Bond series?"). These competency questions were designed to take into account two types of users that could benefit from this ontology schema and the corresponding repositories: *book trade manager*; and *sales director* at an entertainment company. These two types of users belong to one of the companies mentioned in Table 1: Bibliographic Data Services Limited (BDS).[36] This company provides data on books and home entertainment releases, web development and maintenance services, and web-based applications on media to retailers, e-tailers, publishers, libraries, charities and government bodies. BDS offers information to companies about books, audiobooks, music, films and video games, aggregating and extending this data with images, sounds, video clips, screenshots, descriptions, content pages and artist biographies. The DMA ontology schema provides to BDS with a standardised schema that represents their data and allows enrichment with social information, semantic relationships among assets, asset syndication in second screen

[36]https://www.bdslive.com

platforms and, finally, integration with third party datasets.

Considering the focus of the SAM project, the purpose of the competency questions was two-fold: (i) to translate these questions produced by stakeholders in natural language into SPARQL to query the ontology; and (ii) to assess whether ontology repositories based on the DMA schema could provide a correct answer to those questions.

Table 6 shows two examples from the set of 28 competency questions, including the user type, the SPARQL equivalent, and the output obtained after querying the ontology repositories. In addition, the full document containing the competency questions is provided.[37]

[37] https://github.com/knowledge-learning/Digital-Media-Asset.

Table 6: Example of two competence questions for validating the ontology, the translation to SPARQL and the results obtained

Query	Which Martin Campbell films are based on books?
User type	Director of Sales of an entertainment company

SPARQL

```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
PREFIX schema:<https://schema.org/>
PREFIX rdvocab:<http://rdvocab.info/ElementsGr2/>
PREFIX dbo:<http://dbpedia.org/ontology/>
PREFIX dbp:<http://dbpedia.org/property/>
PREFIX dc:<http://purl.org/dc/elements/1.1/>
PREFIX purl:<http://purl.org/ontology/po/>
PREFIX dmaInst:<https://w3id.org/media/dma/JamesBond_CaseStudy#>
PREFIX dma:<https://w3id.org/media/dma#>

SELECT Distinct ?film
WHERE { ?person schema:title ?personTitle.
?film ?contributor ?person.
?film rdf:type schema:Movie.
?film dbo:basedOn ?book.
?book rdf:type schema:Book
FILTER regex(str(?personTitle), 'Martin Campbell')}
```

Result	dmaInst:Casino_Royale
Query	Which authors have written books in the James Bond series?
User type	Book Trade Manager

SPARQL

```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
PREFIX schema:<https://schema.org/>
PREFIX rdvocab:<http://rdvocab.info/ElementsGr2/>
PREFIX dbo:<http://dbpedia.org/ontology/>
PREFIX dbp:<http://dbpedia.org/property/>
PREFIX dc:<http://purl.org/dc/elements/1.1/>
PREFIX purl:<http://purl.org/ontology/po/>
PREFIX dmaInst:<https://w3id.org/media/dma/JamesBond_CaseStudy#>
PREFIX dma:<https://w3id.org/media/dma#>

SELECT Distinct ?author
WHERE { { ?book rdf:type schema:Book.
?book purl:series ?seriesTitle.
?book schema:author ?author.
?author rdf:type schema:Person.
?author rdvocab:professionOrOccupation dmaInst:WRITER.
FILTER regex(str(?seriesTitle), 'James Bond')}
UNION {?book rdf:type schema:Book.
?book purl:series ?seriesTitle.
?book schema:contributor ?author.
?author rdf:type schema:Person.
?autoG owl:annotatedSource ?book.
?autoG owl:annotatedTarget ?author.
?autoG owl:annotatedProperty schema:contributor.
?autoG dbo:role dmaInst:WRITER.
FILTER regex(str(?seriesTitle), 'James Bond')}}
```

Result	dmaInst:Ian_Fleming

The validation of the competency questions was carried out in three stages. In the first one, stakeholders formulated competency questions in natural language and, since they were not SPARQL experts, these were transformed into SPARQL language with the assistance of the ontology experts. Stakeholders knew the DMA ontology schema beforehand, since they contributed actively in its design. Hence, it was not necessary to instruct them on the classes and properties within the ontology before formulating their questions. The transformation from natural language into SPARQL took about two minutes and a half, on average, per competency question. The translated questions were initially verified on an internal repository of the SAM Project. Such repository was populated with 46,228 assets, property of BDS. In the second stage, a second repository[38], focused on the *James Bond* series, was manually developed by processing a sample of digital documents obtained from DBpedia. This repository can be freely queried online by using any query tool (e.g. Protégé) and the competency questions formulated in Table 6 or any other query freely generated by the users.

Considering the DMA schema and the ontology repository, the verification process of the competence questions obtained 100% accuracy (i.e., 28 out of 28 correct answers). This means that the approach followed was reliable enough for extracting personalised information depending on the real users' needs in a real entertainment environment.

Besides that, there was consistency in the DMA ontology instances regarding the needs of the different SAM components. The construction process of the SAM framework guided the development of the DMA ontology schema to ensure compliance with component interoperability requirements. This process gradually gave origin to different DMA properties: the *Extensions* (see Figure 6).

For example, the *Syndicator* component makes use of the syndication extensions to identify which visual widget can represent specific assets and the visual theme to use. This kind of information is represented in the asset as shown below.

Listing 2: Example of a syndication extension in JSON-LD format. JSON attributes are highlighted in bold.

```
"dma:SyndicationExtension": [ {
    "dma:theme": {
        "@id": "dmaIns:THEME1",
```

[38] https://w3id.org/media/dma/JamesBond_CaseStudy, last access March 2019

```
        "@type": ["dma:Theme", "dma:Nomenclature"]
    },
    "dma:widget": {
        "@id": "dmaIns:WIDGET1",
        "@type": ["dma:Widget", "dma:Nomenclature"]
    } } ]
```

This metadata refers to the use of a theme *dmaIns:THEME1* and the widget *dmaIns:WIDGET1* when a specific asset is presented visually in the SAM platform. Notice that the respective code of these elements resides in any internal repository or code library of the platform. The ontology schema represents the interoperability of the components' metadata. This reflects that 100% of the information required by this multimedia platform could be represented.

5. Technical Setup

An API service was developed as part of the *Semantic Services* component to provide a consistent backend service to manage different semantic repositories based on the DMA ontology schema. The *Semantic Services* is a standalone component providing CRUD operations that can be deployed in different frameworks as long as the DMA ontology schema is accessible and there is a semantic server storing instances (see Figure 10).

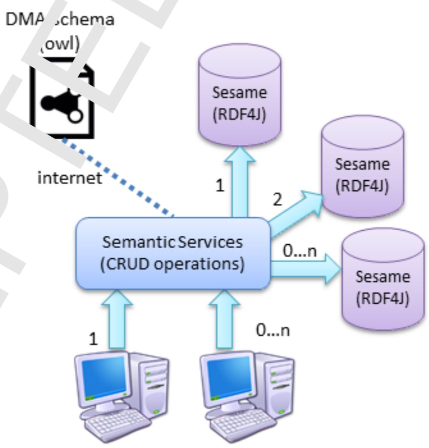

Figure 10: Interoperability of the CRUD services

The aim of the CRUD operations is to provide semantic exploration and management for both assets and nomenclatures. In addition, a service was added for collecting assets related to another asset by using object properties as a query. There are some aspects to bear in mind when using these services:

1. Assets identifiers (*@id*) are automatically generated, avoiding identifier conflicts and allowing unique URIs.
2. An eTAG hash[39] (*__etag__*) is included in every semantic data transfer object (a JSON-LD in this case) to avoid modifications in parallel of the same asset. The eTag hash allows recognising third-party systems requests when managing a specific asset.
3. Every service requires a security token as a parameter, which is used to recognise third-party systems that are allowed to consume and manage the services.

Each CRUD operation includes a RESTful interface and a documentation demo page developed with Swagger[40] (see Figure 11). These RESTful services use JSON-LD objects as semantic inputs and outputs. The parameters required are those that specify attributes to deal with semantic repositories (context, repository name, repository URL, and security token), together with the body parameter that deals with the asset JSON-LD instance. The returned messages provide information about possible responses when the services run. Appendix A shows a complete example of a DMA ontology instance serialised using JSON-LD.

It is important to emphasise that CRUD operations use RDF4J services to guarantee the persistence of assets and nomenclatures in a semantic repository. RDF4J offers an easy-to-use API service that can be connected to RDF4J database servers. The functionalities provided include creating, parsing, storing, reasoning and querying with RDF and linked data. The RDF4J dashboard can be used to exploit the DMA ontology databases, allowing semantic inference to be freely applied.

The *Read* operation service also provides the flexibility of using Construct SPARQL queries.[41] The result of these queries is a single RDF graph formed by taking each query solution in the solution sequence, substituting the variables in the graph template and combining the triplets into a single

[39] https://en.wikipedia.org/wiki/HTTP_ETag.
[40] http://swagger.io/.
[41] https://www.w3.org/TR/rdf-sparql-query/#construct.

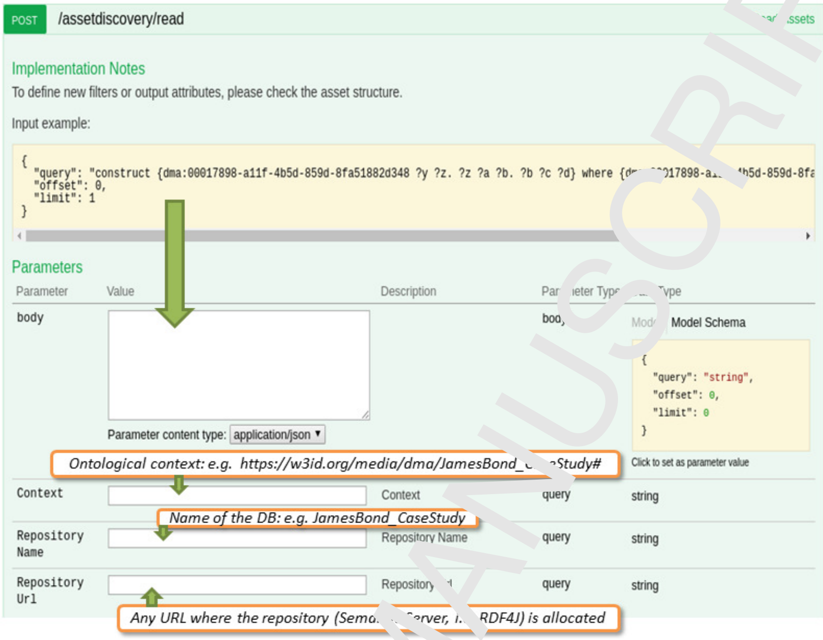

Figure 11: Read service example

RDF graph. To apply these operations, it is necessary to follow the DMA ontology schema described in previous sections.

Other types of CRUD operations are required in the framework defined, such as those regarding asset nomenclatures. An example of a nomenclature instance in JSON-LD is shown below.

Listing 3: Example of a nomenclature in JSON-LD format. JSON attributes are highlighted in bold.

```
{ "@context" : [    "schema":"https://schema.org/",
    "dmaIns":"https://w3id.org/media/dma/
       JamesBond_CaseStudy#"
    },
  "@id":"dmaIns:SPAIN",
  "@type":["schema:Country"] }
```

Regarding the maintainability and sustainability plans, the resources generated through this work are part of the main semantic framework of ongoing and future research projects under the GPLSI[42] (Information Systems and Language Processing Group) of the University of Alicante. Hence, the DMA's sustainability and maintenance is guaranteed. Plans are in place to ensure the support and extension of the resources described in this work to enhance knowledge discovery and representation technologies. To guarantee permanent access to the resources, the DMA repository is located at GitHub[43] and uses the services of permanent identifiers for the Web provided by `https://w3id.org`.

6. Assets Enrichment and Linking: New Opportunities

Besides adding value to existing products, the technological evolution of Artificial Intelligence (which involves knowledge representation), is making it easier for people to get access to, and take advantage of, different kinds of knowledge about everyday tasks [41]. Digital technologies have significantly increased the opportunities for both the entertainment industry and end users. For example, many consumers have already transformed their day-to-day cultural experiences through social media. It is now easier than ever to discover new content and cultural outlets. This may be as simple as searching for a film, book, game or a song recommended online. While these changes are still in their early stages, in many aspects the culture industries have been at the forefront of what is today commonly known as "digital transformation". This transformation refers to various interdependent processes across the digital media asset lifecycle: from importing and aligning media content into a common repository and its semantic annotation, to the creation of asset compositions and their social awareness syndication to end users. All of this can be possible thanks to the consideration of all the features presented in a rich ecosystem such as the one showed in Figure 1.

Content in this ecosystem includes the following features: (i) semantically and socially enabled; (ii) dynamic; and (iii) an efficient schema for the description and representation of media assets. To this end, the DMA schema is extendable and compliant with popular and widely adopted approaches available nowadays. Moreover, it simplifies the digital asset importing process, since standard annotations were reused, allowing links to entities such as Europeana and ontologies from Schema.org, DBpedia, etc. Data from partners or external information resources with references

[42] `https://gplsi.dlsi.ua.es`
[43] `https://github.com/knowledge-learning/Digital-Media-Asset`.

to entertainment entities could be connected to other initiatives and institutions by means of unique identifiers (i.e. URIs), adding and sharing enriched content.

Within the Monitoring and Evaluation (M&E) ecosystem[4], new technologies have exposed the difficulties caused by the current poor state of content metadata collection, curation and standardisation. Through these kinds of technologies, the requirement of Small and Medium Enterprise (SME) content providers is met, which will be able to link to their metadata on demand rather than using complex and expensive data feeds. This will, in turn, maximise sales revenues by providing rich consumer experiences that are intelligent and engaging.

Entertainment companies reveal, through European projects and other collaboration opportunities, their focus on enhancing metadata to engage the consumers by linking products from different categories and by adding rich content to films, music, games and books, such as artist biographies, trivia and quizzes. Through the DMA ontology, this enhancement will be extended to be semantically linked across the M&E ecosystem. For example, assets for a topic such as *James Bond* could be characterised and semantically linked to create a data cluster as shown in Figure 12.

6.1. Use Case

Digital media asset enrichment and linking should guarantee the following features:

- Dynamically discoverable rich assets: achieved by navigating through sematic links (see all semantic relationships of Figure 7 and examples in Figure 13).
- Introduction of social data ratings and reviews, and links to events and/or trends including social data as shown by the object property *dma:socialMediaExtension* in Figure 7.
- Links to other relevant data sources, for example Wikipedia and Getty Images (see in Figure 7 the properties *dma:relatedDbpedia* and *schema:image* respectively).
- Provision of high value contextual data through semantic enrichment.

[4] http://web.undp.org/evaluation/documents/handbook/me-handbook.pdf last access March 2019

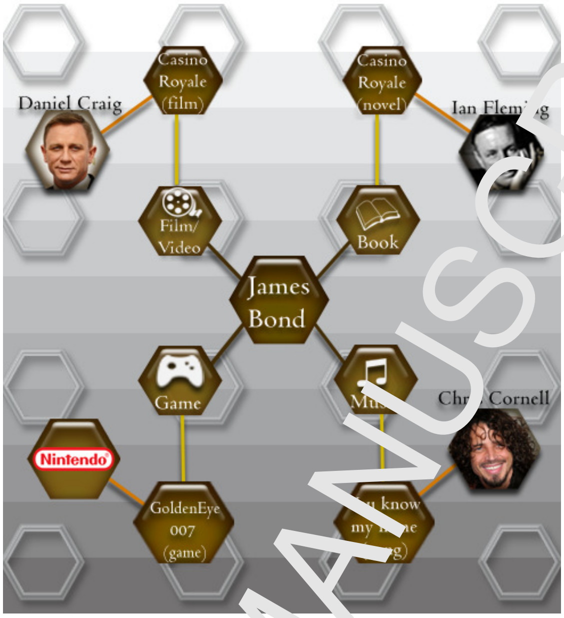

Figure 12: Digital content linking: *James Bond* case study

In the SAM platform, an approach was created to entity linking on two different knowledge Bases. First, the system identifies and links mentions in text to related Wikipedia pages. Secondly, it also identifies references to instances contained in its own media assets knowledge base (e.g. books, songs, films, actors, etc.). These links are created automatically by the *Data Characterisation* component, which belongs to the *Semantic Services* (see Figure 1). The approach, which is further described in [42] and [43], selects possible candidates for each entity mention in text, disambiguating them to both knowledge bases.

Entity linking provides valuable benefits for end users, since they can discover new information about an asset, creating richer experiences around the original content. For instance, a user watching the film *Casino Royale* in the SAM platform could receive information from different knowledge bases thanks to the semantic linking: information related to actors *Daniel Craig* and *Mads Mikkelsen* could be extracted from Wikipedia; and asset instances of a DMA ontology repository could be linked and provided, such as books created by *Ian Fleming*. In addition, social media related sources could be offered, such as the official Facebook page of the *Casino Royale* film and its actors. The previously mentioned *James Bond* case study was developed based on public information provided by DBpedia. This case study can be downloaded and queried by using SPARQL. One of the goals of building this example repository is to demonstrate the potential of the DMA semantic representation for digital data enrichment.

Figure 13 illustrates how different DMA ontology instances can be related, based on semantic relationships (generic and specific). The generic relationships, *relatedAsset* and *relatedDBpedia*, establish links between assets, but do not include a specific role of the asset in the relation. However, more enriched assets could be created supported by more descriptive (specific) semantic relationships. This figure shows how the *relatedAsset* relationship can be refined using other relationships. Due to space limitations, this figure only defines *contributor* (which includes the *role* annotation), *owner*, *sameAs* and *track*, but the whole set of semantic relationships can be explored in the *James Bond* case study example repository. This figure shows how every asset is conceptualised (e.g. *BDS* is an *Organisation* and *Casino Royale novel* is a *Book*) and semantically linked to others. Based on this, entertainment companies could identify how to enrich media content and incorporate standard data transfer objects such as JSON-LD or RDF/XML.

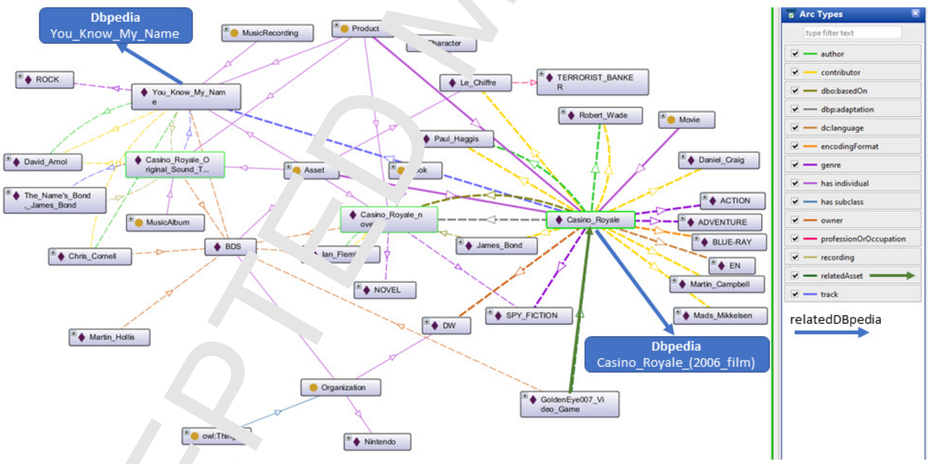

Figure 13: Real scenario simulation

7. Conclusions and Future Work

This research described the DMA ontology schema, designed and developed based on a real scenario. This schema provides features that facilitate the users' engagement with TV programs, films, music and video games, making the information discovery experience more interesting, convenient and personalised.

The construction of this schema has been documented and evaluated following a formal methodology supported by quantitative and qualitative metrics. The results of this work (the ontology schema and an example repository) are freely accessible online. A set of properly tested and documented RESTful services was also developed to manage repositories based on the DMA ontology schema.

The DMA schema and its services have been extended to other research domains beyond the SAM platform. Two other projects have actively supported the further development of this research. These projects are REDES[45] (TIN2015-65136-C2-2-R) and GRE16-01.[46] REDES is a Spanish Government funded project focused on identifying and characterising digital assets by monitoring web content using natural language processing technologies. GRE16-01 is a project funded by the University of Alicante, addressing a framework for text mining to automatically generate semantic knowledge repositories. Both projects are aligned to the DMA ontology schema and use the technologies presented in this paper.

In the future, the plan is to extend the DMA ontology schema to other scenarios to improve the representativeness of diverse digital products that could be semantically related to the current DMA schema classes. In addition, there is a plan to process massive amounts of digital assets retrieved from available database sources, such as DBpedia and Europeana, enriching the instances with the new attributes incorporated in the DMA ontology schema.

Acknowledgment

This research work has been partially funded by the European Commission under the 7th Framework Programme for Research and Technological Development, the Spanish Government, the Generalitat Valenciana and the University of Alicante thought the projects: SAM(FP7-611312); RTI2018-094653-B-C22; RTI2018-094649-B-I00; SIIA(PROMETEU/2018/089); GRE16-01; and Social-Univ 2.0(ENCARGOINTERNO5-19EN).

[45]https://gplsi.dlsi.ua.es/redes/, last access March 2019
[46]https://gplsi.dlsi.ua.es/gplsi13/es/node/396, last access March 2019

References

[1] D. E. M. Association, Ema core metadata summary (for audiovisual digitally delivered content), Tech. rep. (2012).
URL http://www.entmerch.org/digital-ma/resources/5-11-12vid-md-summary.pdf

[2] M. C. Suárez-Figueroa, G. A. Atemezing, O. Corcho, The landscape of multimedia ontologies in the last decade, Multimedia Tools and Applications 62 (2) (2013) 377–399. doi:10.1007/s11042-011-0905-z.
URL https://doi.org/10.1007/s11042-011-0905-z

[3] J. Hunter, Enhancing the semantic interoperability of multimedia through a core ontology, IEEE Transactions on Circuits and Systems for Video Technology 13 (1) (2003) 49–58. doi:10.1109/TCSVT.2002.808088.

[4] A. Miles, B. Matthews, M. Wilson, D. Brickley, Skos core: Simple knowledge organisation for the web, International Conference on Dublin Core and Metadata Applications (2005) 3–10.
URL http://dcpapers.dublincore.org/pubs/article/view/798

[5] R. Arndt, R. Troncy, S. Staab, L. Hardman, COMM: A core ontology for multimediaannotation, in: Handbook on Ontologies, Springer, Berlin, Heidelberg, 2009, pp. 403–421. doi:10.1007/978-3-540-92673-3_18.

[6] S. Borgo, C. Masolo, Foundational Choices in DOLCE, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 361–381. doi:10.1007/978-3-540-92673-3_16.

[7] D. Tarrant, B. O'Steen, T. Brody, S. Hitchcock, N. Jefferies, L. Carr, Using oai-ore to transform digital repositories into interoperable storage and services applications, The Code4Lib Journal 6 (6) (2009) 1–11.
URL http://eprints.soton.ac.uk/267265/

[8] J. Ronallo, Html5 microdata and schema.org, Code4Lib Journal (16) (2012) 1–10.
URL http://journal.code4lib.org/articles/6400

[9] D. U. Board, Dcmi metadata terms (2012).
URL http://dublincore.org/documents/dcmi-terms/

[10] C. Saathoff, A. Scherp, Unlocking the semantics of multimedia presentations in the web with the multimedia metadata ontology, in: Proceedings of the 19th International Conference on World Wide Web,

WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010, 2010, pp. 831–840. doi:10.1145/1772690.1772775.
URL `https://doi.org/10.1145/1772690.1772775`

[11] B. Haslhofer, A. Isaac, data.europeana.eu: The europeana linked open data pilot, International Conference on Dublin Core and Metadata Applications (2011) 94–104.
URL `http://dcpapers.dublincore.org/pubs/article/view/3625`

[12] M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, N. Lindström, Json-ld 1.0: A json-based serialization for linked data, Tech. rep. (2013).
URL `http://json-ld.org/spec/FCGS/json-ld/20130328/`

[13] D. Beckett, T. Berners-Lee, E. Prud'hommeaux, G. Carothers, Rdf 1.1 turtle (2014).
URL `http://www.w3.org/TR/2014/REC-turtle-20140225/`

[14] M. Horridge, P. F. Patel-Schneider, Owl 2 web ontology language manchester syntax (second edition), W3C Working Group Notes (2012) 1–10.
URL `http://www.w3.org/TR/owl2-manchester-syntax/`

[15] K. Coyle, D. Hillmann, Resource description and access (rda) cataloging rules for the 20th century (2007). doi:10.1045/january2007-coyle.
URL `http://www.dlib.org/dlib/january07/coyle/01coyle.html`

[16] F. Gandon, G. Schreiber, Rdf 1.1 xml syntax, Tech. rep. (2016).
URL `http://www.w3.org/TR/2014/REC-rdf-syntax-grammar-20140225/`

[17] I. Hermann, B. Adida, M. Sporny, M. Birbeck, Rdfa 1.1 primer - third edition (2015).
URL `https://www.w3.org/TR/xhtml-rdfa-primer/`

[18] R. Khare, T. Çelik, Microformats: a pragmatic path to the semantic web, in: Proceedings of the 15th International Conference on World Wide Web, ACM, New York, NY, USA, 2006, pp. 865–866. doi:10.1145/1135777.1135917.

[19] IBM, Introduction to business system development method, Tech. Rep. GE19-5387-01, International Business Machines Corporation (1990).

[20] M. Uschold, M. King, Towards a methodology for building ontologies, in: In Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95, 1995.

[21] D. Tansley, C. Hayball, Knowledge-Based Systems Analysis and Design: A Kads Developer's Handbook, British Computer Society practitioner series, Prentice Hall, 1993.

[22] KBSI, Knowledge based systems incorporated, Tech. rep., Wright-Patterson Air Force Base, Ohio (1994).

[23] T. R. Gruber, Toward principles for the design of ontologies used for knowledge sharing?, International Journal of Human-Computer Studies 43 (5) (1995) 907–928. doi:https://doi.org/10.1006/ijhc.1995.1081.

[24] S. Staab, R. Studer, H. Schnurr, Y. Sure, Knowledge processes and ontologies, IEEE Intelligent Systems 16 (1) (2001) 26–34. doi:10.1109/5254.912382.
URL https://doi.org/10.1109/5254.912382

[25] H. S. Pinto, S. Staab, C. Tempich, DILIGENT: towards a fine-grained methodology for distributed, loosely-controlled and evolving engineering of ontologies, in: Proceedings of the 16th Eureopean Conference on Artificial Intelligence, ECAI'2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004, Valencia, Spain, August 22-27, 2004, 2004, pp. 393–397.

[26] A. Gómez-Pérez, M. C. Suárez-Figueroa, Neon methodology for building ontology networks: a scenario-based methodology, in: Proceedings of International Conference on Software, Services & Semantic technologies (S3T 2009), Universidad de Sofia, Bulgaria, 2009.
URL http://oa.upm.es/5475/

[27] M. C. Suárez-Figueroa, A. Gómez-Pérez, M. Fernández-López, The neon methodology framework: A scenario-based methodology for ontology development, Applied Ontology 10 (2) (2015) 107–145. doi:10.3233/AO-150145.

[28] M. Fernández-López, A. Gómez-Pérez, N. Juristo, Methontology: From ontological art towards ontological engineering, in: Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series, American Association for Artificial Intelligence, 1997, pp. 33–40.
URL http://oa.upm.es/5484/

[29] E. Sirin, B. Parsia, Sparql-dl: Sparql query for owl-dl, in: CEUR Workshop Proceedings, Vol. 258, 2007.

[30] D. Tomás, Y. Gutiérrez, I. Moreno, F. Agulló, M. Tiemann, J. V. Vidagany, A. Menychtas, Socialising around media (sam): Dynamic social and media content syndication for second screen, Procesamiento de Lenguaje Natural 55 (2015) 181–184.

[31] D. Tomás, Y. Gutiérrez, Human language technologies in media consumption: The case of sam, Impact of Social Media on TV Content Consumption - New Market Strategies, Scenarios and Trends 3 (2) (2015) 1–7.

[32] J. Abbott, C. Blais, D. Cutts, The standards landscape - siso standards for operations, systems and ontologies (2011).

[33] A. Psychas, A. Menychtas, C. Santzaridou, T. Varvarigou, Y. Gutiérrez, I. Moreno, D. Tomás, Media content linking, semantic annotation and syndication in social enabled multiscreen environments, in: eChallenges e-2015 Conference, Vilnus, Lithuania, 2015, pp. 1–10. doi:10.1109/eCHALLENGES.2015.7441085.

[34] ISO, Iso 9000:2015(en): Quality management systems. fundamentals and vocabulary (2016).
URL `https://www.iso.org/obp/ui{\#}iso:std:iso:9000:ed-4:v1:en`

[35] A. Gómez-Pérez, Ontology evaluation, in: Handbook on Ontologies, Springer, 2004, pp. 251–273.

[36] S. Tartir, I. B. Arpinar, M. Moore, A. P. Sheth, B. Aleman-Meza, Ontoqa: Metric-based ontology quality analysis, IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources (2005) 45–53doi:10.1.1.131.2087.

[37] H. Yao, A. M. Orme, L. Etzkorn, Cohesion metrics for ontology design and application, Journal of Computer Science 1 (1) (2005) 107–113. doi:10.3844/jcssp.2005.107.113.

[38] E. Blomqvist, A. Öhgren, K. Sandkuhl, Ontology construction in an enterprise context: Comparing and evaluating two approaches, in: ICEIS (3), 2006, pp. 86–93.

[39] V. Cross, A. Pal, An ontology analysis tool, International Journal of General Systems 37 (1) (2008) 17–44. doi:10.1080/03081070701499997.

[40] R. Guha, R. McCool, Tap: A semantic web test-bed, Journal of Web Semantics 1 (1) (2003) 81 – 87. doi:https://doi.org/10.1016/j.websem.2003.07.004.
URL http://www.sciencedirect.com/science/article/pii/S1570826803000064

[41] J. Howe, Knowledge-based systems and artificial intelligence: emerging technology, Future Generation Computer Systems 7 (1) (1991) 55 – 68, emerging Information Technologies and Their Impact on the End-user. doi:https://doi.org/10.1016/0167-739X(91)90016-Q.
URL http://www.sciencedirect.com/science/article/pii/0167739X9190016Q

[42] D. Tomás, Y. Gutiérrez, F. Agulló, Entity linking in media content and user comments: Connecting data to wikipedia and other knowledge bases, in: eChallenges e-2015 Conference, 2015, pp. 1–10. doi:10.1109/eCHALLENGES.2015.7441053.

[43] D. Tomás, Y. Gutiérrez, I. Moreno, GPLSI Wikipedia Characterisation v1.0: Entity Discovery and Linking to Wikipedia (2017-06-01).
URL http://hdl.handle.net/10045/66373

Dr. David Tomás is a Lecturer in the Department of Software and Computing Systems at the University of Alicante, Spain. His research interests include information retrieval, knowledge representation, information extraction, question answering, sentiment analysis, recommender systems, log analysis, and machine learning approaches to text categorization. He is the author of more than sixty scientific publications in international conferences and journals. He has participated in over twenty public and private projects, including EU-funded QALL-ME [FP6 IST-033860], FIRST [FP7-287607] and SAM [FP7-611312] projects.

Isabel Moreno holds a degree in Computer Science and a Master's degree in Computer Science technologies, both from the University of Alicante, Spain. She is a technician at the University of Alicante for a Spanish funded project called Canonical Representation and transformations of texts applied to the Human Language Technologies (TIN2015-65100-R). Previously, she participated in EU-public funded projects: A Flexible Interactive Reading Support Tool (FP7-287607) and Dynamic Social & Media Content Syndication for 2nd Screen (FP7-611312). She is also a part-time lecturer at this University and a PhD candidate, her interest is in language and domain independent entity typing systems.

Dr. Yoan Gutiérrez is a lecturer at the University of Alicante, Spain. He has produced more than 60 scientific publications in journals and important international congresses. His research has been mainly focused on the semantic analysis of human language, development of semantic resources and their applications in other tasks such as opinion mining and automatic generation of knowledge. He has been involved in at least six research projects, i.e. Nationals and Europeans (i.e. SAM [FP7-611312]). He has been a member of organizing committees of congresses and scientific committees of several events.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Highlights

- Formalisation and implementation of the Digital Media Asset ontology schema
- Detailed evaluation with different ontology metrics
- Simulation on a real entertainment scenario
- Permanent access to the ontology schema, example scenario and documentation
- API service development for managing semantic repositories using Eclipse RDF4J